# Analisi Pre Processing

ELABORAZIONE DEI DATI SCIENTIFICI

[262-009]

**Autore:**

Testoni Manuel (219155)

November 17, 2025

# Contents

# 1

# Introduzione

Questo studio mira all'analisi sul dataset relativo alle farine animali ponendo un particolare focus sulle classi: "bovino", "pesce" e "pollo".

## 1.1 Obiettivo

L'obiettivo di questa analisi è capire quale, e in che contesti, un determinato pre processing sia meglio di altri. All'interno di questo report sono stati testati le seguenti tecniche di pre processig:

- None: nessun pre processing, proponendo l'analisi sul dataset non processato.
- Normalize: usando il pre processing di normalizzazione
- Baseline
- msc
- Prima derivata
- Seconda derivata

  Una volta effettuata l'analisi andando a processare il dataset tramite tutte queste tecniche andremo a commentare la migliore.

## 1.2 Descrizione del Dataset

Il dataset FarineANIM contiene misure spettrali NIR di farine animali appartenenti a tre categorie: bovino (28 campioni), pesce (26 campioni) e pollo (30 campioni).
Il file `farineanimNIR.mat` include:

- `farineanimNIRdata`, una matrice in cui ogni riga rappresenta uno spettro NIR;
- `category`, il vettore delle classi (pollo, bovino, pesce);
- `assexscale`, il vettore delle lunghezze d'onda da utilizzare come asse x nei grafici e per l'interpretazione dei loadings.

È inoltre disponibile il file `animal_feedNIR.mat`, già formattato come dataset compatibile con il PLS-Toolbox.
Nel report vanno mostrati gli spettri grezzi e quelli preprocessati (colorati per categoria), insieme ai risultati dell'analisi PCA: grafici degli scores, grafici dei loadings come line plot basati su `assexscale` e interpretazione delle regioni spettrali maggiormente utili per la distinzione tra le tre specie.

# 2

# Risultati

In questa sezione vediamo i grafici richiesti per le due migliori tecniche di pre processing.

## 2.1 Spettri Originali

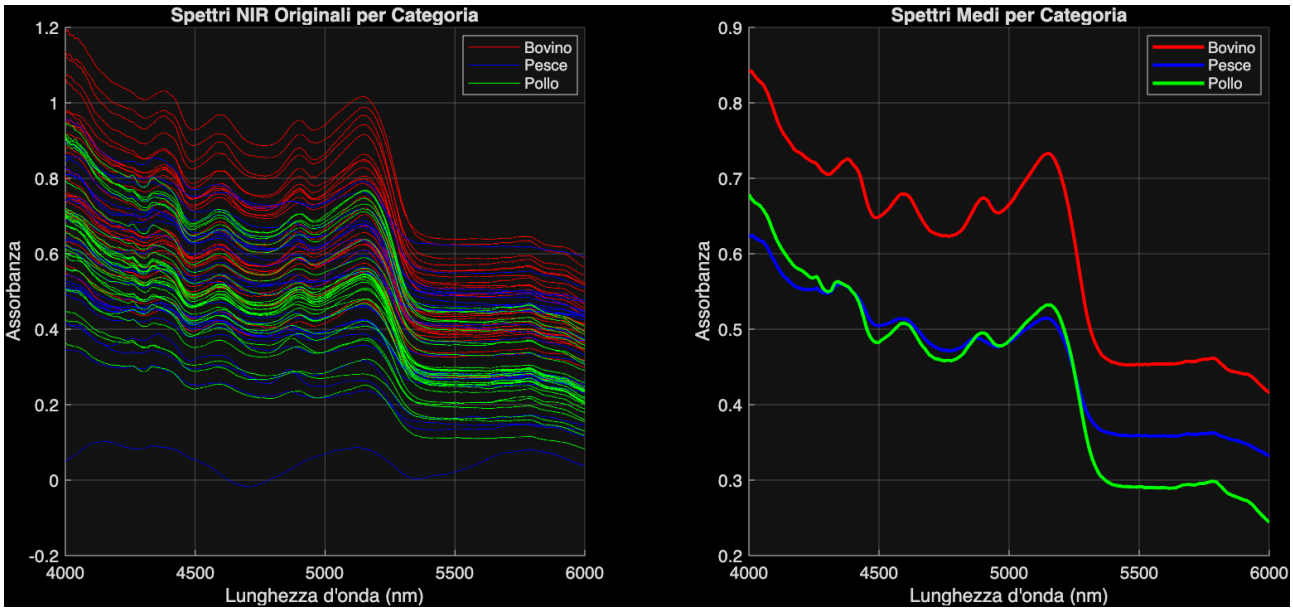


Figure 2.1: Spettri originali del dataset.

Da questi spettri originali vediamo come la classe bovino sia abbastanza separata e su range di valori diversi dalle altre due classi che tendono invece a sovrapporsi. Per differenziare meglio questi dati e per standardizzarli secondo determinati criteri, andiamo, nel corso dello studio ad applicare differenti tecniche di pre processing.

## 2.2 Pre Processing

La scelta del pre proccessing è stata effettuata guardando anche la separazione dei campioni nello spazio PCA effettuato dopo il pre processing. MSC, SNV e Normalizzazione sono stati scartati perché, pur riducendo effetti di scattering, normalizzano eccessivamente gli spettri rimuovendo informazioni discriminanti legate all'intensità assoluta. Questo ha causato una drastica riduzione della varianza tra categorie, rendendo impossibile la separazione nei plot PC1-PC2. La proiezione dei punti nello spazio PCA effettuata dopo la tecnica di pre processing "Baseline" ha prodotto campioni meno sovrapposti rispetto alle 3 tecniche elencate sopra, ma comunque molto ravvicinati. Senza applicare alcun pre processing vi erano aree in cui era possibile definire che quei punti erano appartenenti ad una classe, ma vi erano anche aree sovrapposte. La seconda derivata aveva ottenuto una separazione migliore rispetto alla prima andando però ad introdurre nuovamente rumore. Dunque il pre processing che ha soddisfatto i criteri di scelta è stata la **derivata prima**.

## 2.3 Grafici

### 2.3.1 Spettri Derivata Prima

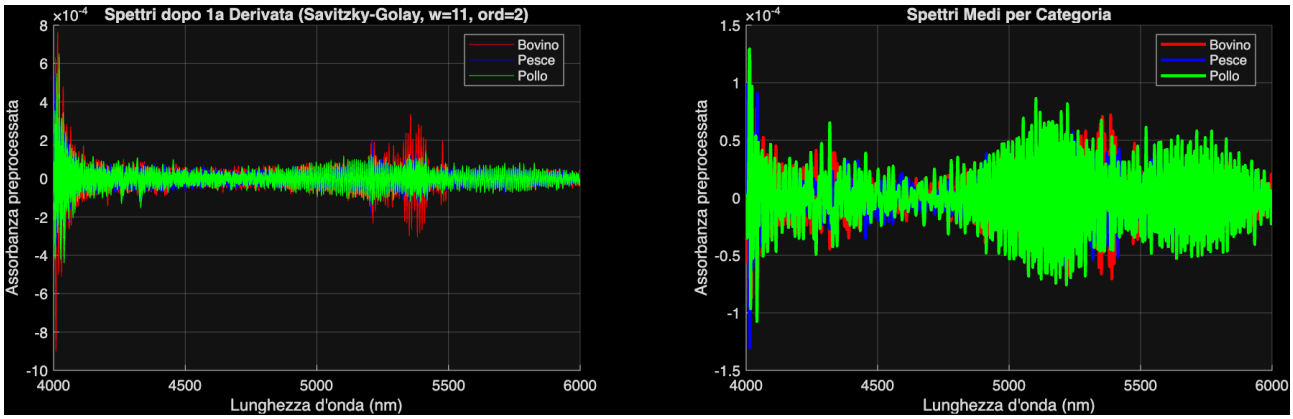


Figure 2.2: Spettri derivata prima.

Questa foto rappresenta gli spettri dopo applicazione della prima derivata (Savitzky-Golay, finestra=11, ordine=2). Il preprocessing rimuove il baseline drift ed enfatizza le differenze spettrali tra le categorie. Da questo spettro riusciamo a vedere come la classe "pesce" venga quasi per tutto lo spettro coperta dalla classe "pollo" mentre la classe "bovino" in qualche punto dello spettro riesce a essere individuata.

## 2.4 Analisi PCA

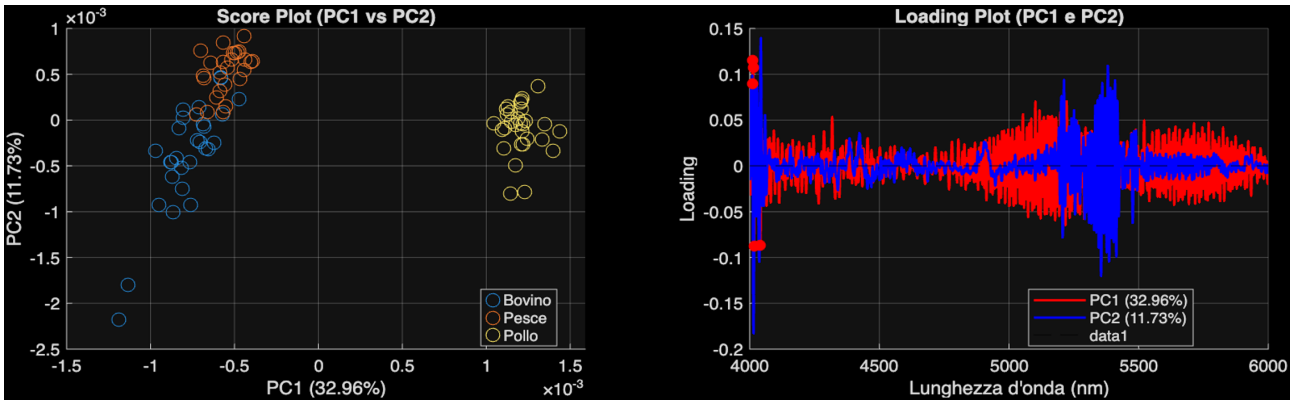


Figure 2.3: Scores per le componenti 1 e 2 di PCA.

Dal grafico in figura siamo in grado di vedere gli score plot (PC1 vs PC2) dopo 1a derivata. Si osserva una netta separazione tra le tre categorie di farine. PC1 spiega 34% della varianza, PC2 spiega 13%. In questo grafico notiamo come la categoria "pollo" risulta separarsi completamente dalle altre nello spazio PCA, mentre Bovino e pesce hanno un punto dove la separazione è meno evidente ma nel complesso sono ben separate.

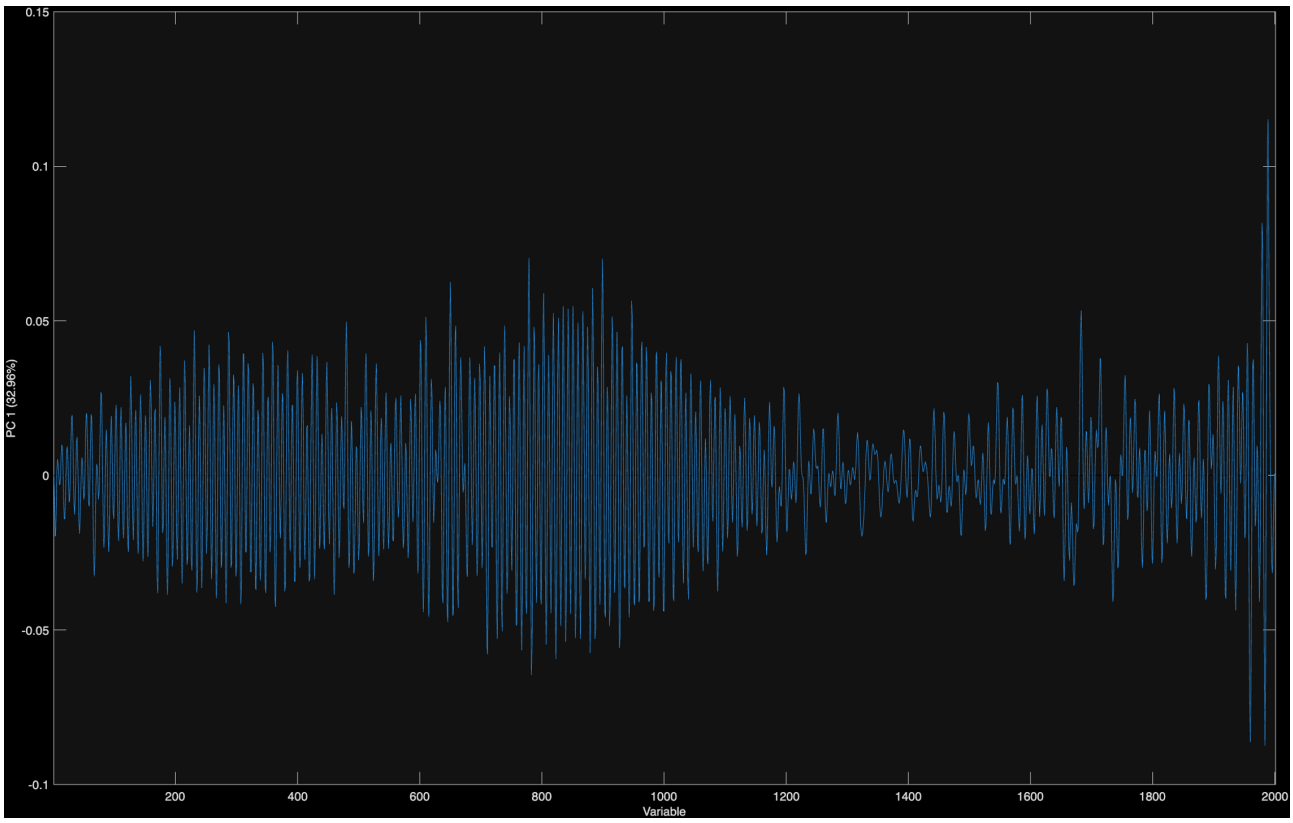


Figure 2.4: Spettro dei loadings.

Da questa figura possiamo determinare 3 cose:

- Picco massimo a 5650-5770 nm (variabile 1750-2000): Assegnazione: C-H stretch 1 e 2 sovratono dei lipidi (gruppi $CH_2$) Interpretazione: Discrimina in base al contenuto di acidi grassi saturi (bovino/pollo) vs insaturi (pesce) Regione 4800-5300 nm (variabile 700-1000):

- Assegnazioni: N-H stretch proteine (4878-5051 nm, amide II) e O-H stretch + C-O amido/acqua (4762-5263 nm) Interpretazione: Rapporto proteine/carboidrati differenzia bovino (più proteico) da pollo (presenza cereali) Regione 6300-6600 nm (variabile 500-700):

- Assegnazioni: N-H stretch proteine (6623 nm) e O-H stretch amido/glucosio (6329-6545 nm) Interpretazione: Contenuto proteico vs carboidrati complessi

  Come si è riusciti a ottenere queste analisi? Facendo riferimento al file "tointerpretNIR" e utilizzando la seguente formula per convertire l'indice della variabile $i$ nel loading plot alla corrispondente lunghezza d'onda $\lambda$ (nm):

$$\lambda(i) = 4000 + (i-1) \quad \text{con } i = 1, 2, \ldots, 2001 \tag{2.1}$$

dove il dataset contiene 2001 variabili distribuite uniformemente tra 4000 e 6000 nm. Ad esempio, la variabile 1001 corrisponde a 5000 nm, la variabile 1501 a 5500 nm.

## 2.5 Scree Plot

Attraverso questo grafico siamo in grado di capire quante e quali componenti PCA create. Inoltre possiamo anche capire quanta percentuale di varianza si sta effettivamente mantenendo riducendo allo spazio delle componenti principali. Questo ci aiuta nel capire quando è effettivamente il caso di utilizzare PCA e quando no perchè magari stiamo perdendo troppa informazione.

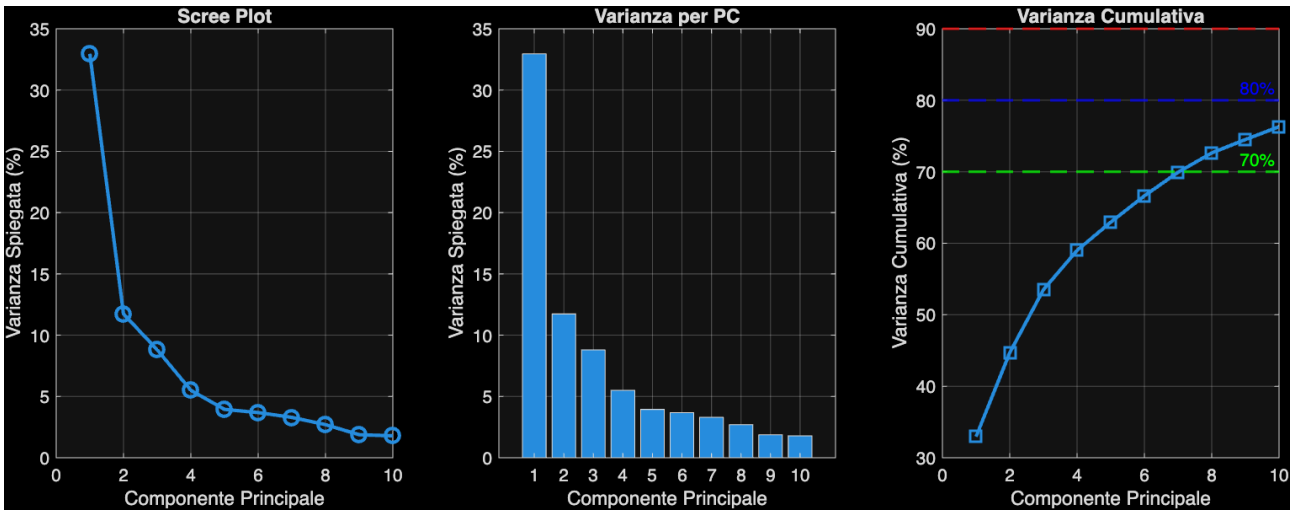


Figure 2.5: Scree plot derivata prima.

Si è dunque scelto di procedere con l'analisi utilizzando due componenti principali in quanto catturano il 48% della varianza nel dataset e anche perchè dopo la seconda componente principapale vi è una sorta di "gomito", ovvero vi è una particolare situazione che va a far sì che pur aumentando significativamente il nuemero di PC non si aumenti proporzionalmente la varianza spiegata. Questo fa sì che servano 8 PC's per spiegare il 70% della varianza all'interno del dataset e dato che questa è una tecnica di decomposizione, ossia di riduzione della dimensionalità dello spazio dei dati, andare ad utilizzare 8 PC's andrebbe a perdere il senso per cui si procede con questa analisi.

# 3

# Conclusioni

In conclusione, sono stati testati 6 diversi metodi di preprocessing sugli spettri NIR delle farine animali. L'analisi ha dimostrato che la 1a derivata prima si è rivelata la tecnica più efficace per la separazione delle tre categorie (bovino, pesce, pollo), fornendo score plot con gruppi ben distinti. L'analisi PCA condotta su 2 componenti principali (48% circa di varianza cumulativa) ha permesso una riduzione dimensionale efficace. L'interpretazione dei loading ha confermato che la separazione è basata sulle differenze composizionali nei lipidi (5650-5770 nm), proteine (4878-5051 nm) e carboidrati (4700-5300 nm), in accordo con le caratteristiche biochimiche note di questi alimenti.